AI answer evaluation and acceptance checks

Prepared for: [CLIENT / BUYER / ROLE]
Prepared by: Suede AI
Version: [VERSION]
Issue date: [DATE]
Status: Editable proposal template; tailor to the named buyer and approve client-specific scope and terms before issue.

Executive summary

[CLIENT] needs to decide whether answers from [AI SYSTEM] preserve [BUSINESS-CRITICAL FACTS] before [RELEASE OR OPERATING DECISION]. [BUYER] has identified [KNOWN FAILURE OR QUALITY GAP]. Suede AI will turn approved sources into a reviewable rubric, build a bounded set of test cases, and provide repeatable checks with failure categories and evidence. Your team can use the results to approve a release, investigate failures and compare later changes on the same agreed cases.

Investment and approval

Complete the fields below for this client. Confirm scope, dates and terms with the client before issuing the proposal.

Field	Client-specific entry	Approval
Engagement fee and currency	[FEE] [CURRENCY]	Client-specific approval required before issue.
Payment schedule and taxes	[PAYMENT MILESTONES; TAX TREATMENT]	Client-specific commercial approval required.
Third-party usage budget	[APPROVED PROVIDERS; SPEND CAP; BILLING OWNER]	Client approval required before paid calls or subscriptions.
Proposed start and delivery dates	[START DATE] / [REVIEW DATE] / [DELIVERY DATE]	Confirm after access, inputs and owners are agreed.
Review window and correction process	[REVIEW WINDOW] / [CORRECTION WINDOW]	Client-specific acceptance procedure requires approval.
Agreement and procurement	[SOW/AGREEMENT REFERENCE; AUTHORISED SIGNATORIES; PROCUREMENT STEPS]	Commercial owner and client approve the governing agreement.
IP, confidentiality and data terms	[OWNERSHIP/LICENCE; CONFIDENTIALITY; DATA PROCESSING; RETENTION; DELETION]	Confirm with the client and accountable legal/privacy owners.
Validity, cancellation and support	[PROPOSAL VALIDITY; CANCELLATION; POST-DELIVERY SUPPORT]	Client-specific terms require approval.

Proposed scope

Evaluation design and a runnable acceptance-check package for [ANSWER TYPE], [APPROVED SOURCES], [CASE SET], [MODEL/SYSTEM VERSION] and [RELEASE DECISION]. Confirm case coverage, scoring rules and model-call budget before execution.

Discovery inputs

- Name the product sponsor, engineering owner, subject expert and person accountable for accept/review decisions.
- Provide approved references with versions and dates, representative questions, known failures and examples of acceptable answers.
- Define the claims an answer must preserve, unsupported claims it must flag, source-link requirements and actions needing human review.
- Agree the target system/model, prompt template, retrieval settings, relevant generation settings, environments and approved data classes.
- Choose the case coverage, failure severity, decision thresholds, review process and cost ceiling for actual model runs.

Deliverables and acceptance

Source-backed rubric

Translate approved facts into named checks, severity levels and accept/review rules. Separate deterministic checks from judgments requiring a subject expert.

Acceptance: The client subject owner approves the reference facts and can explain how each check affects the release decision.

Versioned test cases and fixtures

Create representative cases and controlled failure fixtures with expected outcomes, provenance and review notes. Reserve held-out cases where the agreed sample supports them.

Acceptance: Engineering and subject owners review coverage; controlled failures exercise the named categories and expected decisions.

Repeatable evaluation command

Package a documented runner, input format and report with case results, check outcomes and failure categories. Retain model/system version, prompt, settings and raw output for approved model runs.

Acceptance: The client runs the documented command in the agreed environment and obtains the expected fixture decisions; actual model results link to the original output and run metadata.

Findings and operating handover

Explain observed failures, prioritize remediation, record unresolved judgments and provide the comparison method for a later system change.

Acceptance: Product owner receives the report and can trace a release decision to the approved rubric, case evidence and accountable reviewer.

Phased delivery

The schedule fields are proposed. Client-specific approval is required before dates become commitments.

Define the decision

[PROPOSED RUBRIC WINDOW; confirm with client]

Select the answer type, sources, case coverage and accept/review rules with the business and engineering owners.

Exit criterion: Subject owner approves reference facts and severity; sponsor approves the bounded evaluation scope.

Build and validate checks

[PROPOSED BUILD WINDOW; confirm with client]

Prepare fixtures, implement checks and run controlled cases to verify failure detection and reporting.

Exit criterion: The agreed fixture outcomes match the documented rules and the client can rerun the command.

Evaluate and hand over

[PROPOSED RUN/REVIEW WINDOW; confirm with client]

Evaluate the approved system within the authorized call budget, review judgment-based cases and deliver evidence and remediation priorities.

Exit criterion: The report includes run metadata, raw-output references and reviewer decisions for the agreed cases; product owner records the release or remediation decision.

Responsibilities

- Suede AI: Design the rubric, prepare agreed cases and runner, retain approved run evidence and explain observed failures.
- Client subject/product owner: Approve reference facts and failure severity, review judgment-based cases and own the release decision.
- Client engineering/security owner: Provide the agreed environment and scoped access, approve provider routing and budget, and identify versioned prompts and system settings.

Access and data handling

- Classify the test data and use synthetic or redacted fixtures before handling production examples.
- Keep reference documents, prompts and outputs in the approved workspace. Treat retrieved content and answer text as untrusted input.
- Approve model providers, retention, regional routing and data-processing requirements before an evaluation sends client data.
- Use scoped test credentials and a recorded cost ceiling. The runner should expose the case, source and settings without including secrets in logs.
- Agree who can review sensitive outputs, where evidence is stored and when raw captures and access are deleted or revoked.

Assumptions and dependencies

- The client supplies authoritative, dated references and a subject owner who can resolve ambiguous cases.
- The agreed case set defines coverage. Results are reported with the denominator, run conditions and separate failure categories.
- Fixture checks validate the evaluation behavior. Model-performance claims require an actual authorized model run with retained prompts, settings, dates and outputs.
- Prompt changes, retrieval redesign, fine-tuning, production rollout and ongoing monitoring are separately approved implementation scopes.

Change control

Either party can request a change in writing. Suede AI records the requested output, affected systems, acceptance checks, cost and delivery impact. The client sponsor and Suede AI commercial owner approve the revision before that work starts. The current approved scope continues while the change is reviewed.

Relevant proof and personal credentials

Four authored answer fixtures produced four expected decisions against eight checks per fixture. Failures covered conflated award totals, a missing source and training rewritten as partnership. [source: authored answer-fixture demonstration results, 2026-10-21]

Proof layer: Deterministic fixture test; use a separately recorded model run for model-performance measurement.

Inspect this working example · Read the dated result

These records belong to Jason Colapietro. Each item retains the issuer's award type. Course badges, academy accreditation, training module badges and learning path trophies describe personal training; company relationships have their own evidence records.

- IBM SkillsBuild: Craft Precise Prompts for AI Models. Digital credential; issued 2026-10-21. Issuer verification [source: issued credential inventory, verified 2026-10-21]
- Google Cloud / Google Skills: Introduction to Large Language Models. Course completion badge; issued 2026-10-21. Issued award artifact [source: issued credential inventory, verified 2026-10-21]
- Google Cloud / Google Skills: Introduction to Generative AI. Course completion badge; issued 2026-10-21. Issued award artifact [source: issued credential inventory, verified 2026-10-21]
- Databricks Academy: Academy Accreditation - Generative AI Fundamentals. Academy accreditation badge; issued 2026-10-21. Issuer verification [source: issued credential inventory, verified 2026-10-21]

Questions for the buying team

Can this compare two system versions?

An approved comparison can run the same versioned cases against each system, preserve the settings and source versions, and report paired outcomes by failure category.

Who decides whether an answer is acceptable?

The client approves the rubric and owns the release decision. Deterministic checks report factual failures; named subject reviewers resolve the judgments defined in the scope.

Next step

Bring [ANSWER TYPE], its approved references and a known failure. Confirm the release decision, case coverage, review owner, model-call budget and client-specific investment before evaluation begins.

Book a scoping conversation · Email Suede AI · +1 (561) 401-0518

Reference sources

- Suede AI operations
- Suede AI qualifications and working proof
- Issued personal training records
- Recorded demonstration results
- Recorded examples and inspection guide